Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 12:47:00
REFERENCE NO: 238

Author Names & Affiliations

- John Towns - NCSA/University of Illinois
- Nancy Wilkins-Diehr - SDSC/University of California, San Diego
- Gregory Peterson - NICS/University of Tennessee
- Ralph Roskies - PSC/University of Pittsburgh
- Kelly Gaither - TACC/University of Texas at Austin
- Dave Hart - NCAR
- Dave Lifka - CRC/Cornell University
- Ron Payne - NCSA/University of Illinois
- Dan Stanzione - TACC/University of Texas at Austin

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

All fields

Title of Submission

Harmonizing Critical Elements of the CI Ecosystem: people, resources, services, software, research

Abstract (maximum ~200 words).

Significant progress has been made in the development of cyberinfrastructure capabilities in recent years. The community is now faced with further advances in capabilities along with a growing need to bring greater harmony to those capabilities to enable researchers to easily and effectively harness those capabilities. This submission provides input from the XSEDE project based on the experiences of the team participating in that project that spans many disciplines and include some who have been participating in the emergence of cyberinfrastructure for more than 30 years.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

In this response, XSEDE will not attempt to enumerate the many research challenges it is exposed to across the breadth of traditional and new disciplines it supports. The focus of this response is to relate needs and other issues based on an integrated view of the significant cross section of the research community with which XSEDE interacts.

XSEDE interacts continuously with users and sees a number of common challenges in accomplishing their research. Most notably, the shortage of production-scale computing, storage and related resources, including computational science and cyberinfrastructure experts, to meet the demand from current agency-funded, merit-reviewed science and engineering awards. In any planning for cyberinfrastructure, the desire for tomorrow’s innovative cyberinfrastructure technologies must be balanced against the need and demand for available, best-of-breed cyberinfrastructure capabilities and capacities for today’s research needs.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Advancing science across multiple disciplines requires a variety of resources and services. There is thus the need for comprehensive cyberinfrastructure composed of heterogeneous digital resources leveraging the aggregate expertise of a small number of leading institutions, each with its own unique human talent and approaches to addressing community needs. However, even with heterogeneous resources and the expertise of multiple institutions, it is only through tight yet flexible integration and interoperability of these resources and services that a growing number of scientific research activities can move forward efficiently and effectively.

Further, there needs to be greater harmonization of the ecosystem among the various providers. While XSEDE provides a significant amount of this, it is still limited in its scope. There are options in how this might be realized, but it is unlikely that a single organization can provide for the needs of all NSF researchers and researchers funded by other agencies that project such as XSEDE are expected to support. With growing demands and shrinking budgets, the nation must find means by which researchers can readily gain access to the myriad of resources without having to stand them all up, support them, etc.

To improve user productivity and thus generate more science and scholarship, researchers need to be presented a single interface rather than a set of different interfaces with different administrative domains. Examples include a single user portal, to coordinate and unify services and information such as training offerings, allocations, documentation, publication management, and the status and performance of resources. Other common services include a single help desk, unified authentication mechanisms, an application software catalog, a repository of services and tools, coordinated security (in order to prevent the spread of security breaches from site to site), unified allocation of resources (which helps assure that the most meritorious work nationally is awarded resources), coordinated advanced support and training, common techniques for rapid data transfer, support for end-to-end network tuning to optimize data transfer, and a well-defined infrastructure that will allow campuses to design their systems to interoperate with others in the broader cyberinfrastructure ecosystem.

Here we give a few examples of the challenging projects scientists expect to carry out over the next five years.

In many fields new to HPC, such as digital humanities, machine learning, genomics, and radio astronomy, there is a trend toward assembling numerous, rapidly-evolving software packages into powerful applications and workflows. For example, researchers nationwide are using the NSF-funded single-dish Green Bank Telescope (GBT) at the National Radio Astronomy Observatory to search for millisecond pulsars, to detect and study gravitational waves, and to study astrochemistry. A project for rapidly spinning neutron stars generates ~1 PB of data per year. The Python-based GBT Mapping Pipeline, which integrates many tools, is a new software system intended to facilitate the production of sky maps from this massive data stream. Assembling and optimizing this pipeline, must maintain high reliability and throughput to keep pace with observations.

Many computational chemists use multiple techniques (quantum simulations, molecular dynamics, QM/MM), optimized on different resources. Researchers need support to develop solutions for these complex workflows. For example, Kendall Houk’s group at UCLA has been studying ruthenium catalysts for olefin metathesis (controlled formation of carbon-carbon double bonds). Highly accurate quantum simulations required more than 100 GB of scratch space per node. On newer systems, these calculations can be done more efficiently entirely in core, allowing for the design of even more complex catalysts.

Two large biology user communities, the Cyverse project(formerly iPlant Collaborative) and the Galaxy Project, encompass more than 35,000 U.S.-based users and operate sizable software infrastructure projects. Experimental biologists would prefer to use these interactively. iPlant and Galaxy Main are currently delivered from an oversubscribed hardware environment not leveraging national resources. Such communities seed the technical support to move these systems to new resources to enable this work to be efficiently carried out, tuning applications and implementing science gateways.

The National Snow and Ice Data Center (NSIDC) curates and manages widely used data but has no community collection of analysis routines. Thus, a polar researcher might know where to get data but not where to find best-practice analysis routines. Moreover, no common computing infrastructure is available to this community. At least 2,500 researchers regularly use NSIDC-managed data products. NSIDC staff do not have the expertise to create and publish virtual machines/containers capable of requesting NSIDC data and running common earth science/polar science routines to enable more effective research and better analyses of data, including automated recording of provenance and version information.

Advanced computing and large-scale data analytics are transforming the way scholars address literature and art. In literature, going beyond the traditional, and arguably limited, approach of close reading, where a researcher carefully analyzes a relatively small body of work, distant reading or macroanalysis applies techniques from natural language processing, statistics, graph analytics, and machine learning to analyze significantly larger corpora, yielding important insights for contextualizing literary movements. The scaling is not merely of the number of works being addressed. Rather than working independently, interdisciplinary and geographically distributed communities of researchers with common interests and complementary expertise are collaborating, building sophisticated infrastructure such as the Collaborative for Historical Information and Analysis (http://www.chia.pitt.edu/) and the Digital Mitford Archive (http://www.digitialmitford.org). This community needs access to such data collections and support in identifying and installing appropriate, scalable applications and frameworks for collaborative, data-intensive research (e.g. collaborative editing, data integration and fusion, GIS and overlay systems, and digitization of scanned works).

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

The progress of science requires a pipeline of people knowledgeable in exploiting the available technologies. Access to advanced technical staff is a critical element to an effective cyberinfrastructure environment. The resources available to researchers change continually, constantly requiring training and education in new techniques and the development of new algorithms.

The centralized coordination and management of these efforts enables the most appropriate experts to be brought to bear to assist users, no matter where the experts or the resources are located. This also provides important stability to the career paths of the valuable experts. For example, the pooling of expertise has allowed ECSS to support experts in digital humanities and workflows, which it is unlikely that an individual center could have provided due to insufficient demand at any one site. That centralization also enables cross-pollination of knowledge between disciplines and resources. We have often seen advanced user support professionals transmitting advances and insights at the algorithmic, numerical, coding, and optimization levels between fields of application and between computing systems.

The development of the workforce is becoming a critical issue. As the subject of a number of recent workshops and other activities, it is become apparent in the community that it is necessary not only to foster the professionalization of the CI workforce, but also to put significant effort into developing programs at universities and colleges that will produce the future workforce that will be needed. This will require studies to assess the projected needs to understand the demand that can then be used to help drive development and instantiation of the necessary educational programs to produce this workforce. The anticipated workforce needs are not well defined, but it is clear that in related fields, the workforce needs will far outstrip the available workforce over the next few years. To be clear, this is far more than systems administrators and network engineers. There is a range of critical roles ranging from those operating hardware to those providing advanced support to research teams and software efforts. There are indicators, identified via the various workshops, of the difficulty in finding qualified candidates for positions today. As campus cyberinfrastructure proliferates, the workforce needed will grow dramatically.

NSF has done an excellent job of developing policy about, and procedures for, allocating compute resources for the national community. It has not done an analogous job when it comes to storage resources. The community needs a coherent policy from NSF regarding long term storage of data. Issues include who operates the storage, and who decides what data is stored and for how long. To date, there is no simple accepted mechanism for a PI to request funding for storage of data for the duration of a grant, and more importantly, no way to assure continued storage after the expiration of the grant. NSF should be developing plans for national facilities, with long lifetimes, that specialize in long term storage of data. And just as the community reviews applications for computer time, there needs to be a way to review proposals for storing data, and for how long. These could be revisited every year for each stored dataset. Special funds might need to be set aside for data storage, with committees representing particular disciplinary communities, rather than individual PIs, deciding which datasets are worth preserving for the next period

For both compute and storage resources, NSF should be developing five-year roadmaps, to be revisited and updated each year, outlining its expected financial investments. This is a critical piece that will allow researchers and research communities to develop their research plans--many of which require years to accomplish and are complicated currently by short-term plans with respect to resources that have seen unexpected changes from the perspective of the community.

XSEDE staff regularly are faced with questions regarding long-term data management and data preservation issues. Often these are related to data management plans associated with research proposals and also the more recent open data access policy of NSF. We suggest that NSF sponsor a workshop to develop recommendations on how NSF should (or perhaps should not) invest in this as part of the NSF strategic roadmap. Particularly difficult issues encountered of late that the community is sorely in need of either solutions or guidance include:
+ What data should be preserved? How should decisions about what data should be preserved be made?
+ Given that grant funds cannot be used to support preservation of data past the end of the award period, how do the infrastructure and services necessary get paid for?
+ How should the open data access directive from the Office of Science and Technology Policy be interpreted and translated into policies and practices for data produced by NSF-supported research?

Consent Statement